63-3-2

(1)

(5) 161 500



*Western Management Science Institute*
*University of California ● Los Angeles*



$1.10

UNIVERSITY OF CALIFORNIA LOS ANGELES

GRADUATE SCHOOL OF BUSINESS ADMINISTRATION

WESTERN MANAGEMENT SCIENCE INSTITUTE

December 17, 1962.

Western Management Science Institute

Working Paper No. 26

# CONVERGENCE PROPERTIES OF AN ALGORITHM FOR LEARNING TO CLASSIFY,

By

Zivia S. Wurtele,

The Author is grateful to Professors T. Ferguson and J. MacQueen for discussions of this paper. Comments are welcomed.

Acknowledgements. This work was supported partly by the Office of Naval Research under Task 047-041, and partly by the Western Management Science Institute under a grant from the Ford Foundation. Reproduction in whole or in part is permitted for any purpose of the United States Government.

# CONVERGENCE PROPERTIES OF AN ALGORITHM FOR LEARNING TO CLASSIFY

By
Zivia S. Wurtele

~~1. Introduction. In this paper I shall discuss~~ a problem which involves a fairly simple form of decision making -- classifying -- and a special type of algorithm for learning to solve it is discussed. The problem is that of classifying correctly individuals which are drawn at random from a population which is partitioned into a finite number of categories. The learning process is required to be a step-by-step procedure in which observations are made on individuals one at a time, and the current estimate of the required partitioning may be adjusted after each observation, on the basis of knowledge of the category in which the individual observed falls. At any given time, the current estimate of the partitioning is all that is held in memory; past history is lost except insofar as it has been incorporated into the present estimate. The learning process of perceptrons, as well as that of other artificial intelligences, is of this general form.

Each individual is characterized by a vector in n-dimensional Euclidean space. I shall assume that this characterization is sufficiently rich with respect to the given classification problem, by which I mean the following. If $S_i$ is the smallest closed convex set which contains all the vectors which describe individuals of the i category, then the intersection $S_i \cap S_j$ of any two such convex sets is empty. This terminology is appropriate to situations for which in the case of failure of the condition of sufficient richness, a re-examination of the world of individuals and the subsequent increasing of the number

of components of the characterizing vectors can be expected to yield a new characterization for which this condition is satisfied. The question of whether, in a particular case, a sufficiently rich characterization can be achieved is obviously crucial but beyond the scope of this paper. Many problems are ruled out by this requirement, including those for which the noise level of the measuring instruments is too high or for which the very act of taking the measurement changes the category of the individual, as well as those which involve relations which are essentially non-linear.

2. Notation and Assumptions. I shall follow the convention of using upper case letters to denote vectors or sets of vectors and lower case letters to denote scalers. In the argument below, each individual, which is characterized by a vector X in n-dimensional Euclidean space, is a member of one and only one of two categories. The results obtained are applicable to the general case, however, for they may be applied to appropriate partitions of a set of three or more categories into two subsets. I make the following assumptions:

(1) $X \geq 0$.

(2) $0 < h_1 \leq |X| \leq h_2 < \infty$.

(3) There exists a plane $B^*X = 1$ and a positive number $c^*$ such that if $X \in S_1$, $B^*X \geq 1 + c^*$ and if $X \in S_2$, $B^*X \leq 1 - c^*$ (Obviously, if one such plane exists, so do an infinity of planes.)

The problem is to estimate a vector B* sequentially so that each time a vector X is observed, the current estimate of B* is subject to revision in accordance with a rule which depends upon knowledge of whether the vector X is in $S_1$ or in $S_2$. This rule will be described in Section 3; and its convergence properties will be discussed in Section 4.

The case where the dividing plane passes through the origin and the vectors are binary was analyzed by Papert [1]; and the algorithm in Section 3 is a natural extension of the one is [1]. The results in this paper are relevant to the case where it is not known that a plane which separates the regions passes through a particular point. If such a point is known, then a translation which moves it to the origin will allow use of the algorithm in [1].

3. The Algorithm. It is assumed that sampling is random and that initially there are two samples: $X_1^1, \ldots, X_1^p$ from $S_1$ and $X_2^1, \ldots, X_2^q$ from $S_2$. Let $X^t$ be the $t^{th}$ vector sampled after the initial $p + q$ vectors. Estimate B* as follows:

(1.) Let the initial estimate of B* be $B^1$, any vector for which the plane $B^1X = 1$ separates the initial $p + q$ vectors so that $B_2^1 X_1^d > 1$, for $d = 1, \ldots, p$, and $B^1X_2^d < 1$, for $d = 1, \ldots, q$.

(2.) Obtain the $(t + 1)^{st}$ estimate from the $t^{th}$ estimate from the equation:

$B^{t+1} = B^t + e_1^t X^t - e_2^t X^t$, where the e's are determined

in accordance with the following rules:

(a) If $X^t \in S_1$, then

$$e_2^t = 0;$$

$$e_1^t = 0, \text{ if } B^t X^t > 1;$$

and $$e_1^t = \frac{(1 - B^t X^t)}{|X^t|^2} + o\left(\frac{1}{t^2}\right), \text{ if } B^t X^t \leq 1.$$

(b) If $X^t \in S_2$, then

$$e_1^t = 0;$$

$$e_2^t = 0, \text{ if } B^t X^t < 1;$$

and $$e_2^t = \frac{(B^t X^t - 1)}{|X^t|^2} + o\left(\frac{1}{t^2}\right), \text{ if } B^t X^t \geq 1.$$

(c) $o\left(\frac{1}{t^2}\right)$ is positive and sufficiently small so that the addition (case a) or subtraction (case b) of the term $o\left(\frac{1}{t^2}\right) X^t$ does not change the sign of any component of $B^{t+1}$.

4. Convergence Properties of the Algorithm. The vector $B^t$ may be written:

$$B^t = B^1 + (e_1^1 X^1 + \ldots e_1^{t-1} X^{t-1}) - (e_2^1 X^1 + \ldots e_2^{t-1} X^{t-1})$$

$$= B^1 + r_1^t Z_1^t - r_2^t Z_2^t,$$

where $$Z_i^t = (e_i^1 X^1 + \ldots e_i^{t-1} X^{t-1}) / r_i^t$$

and $$r_i^t = e_i^1 + \ldots e_i^{t-1}, \text{ for } i = 1, 2.$$

Obviously, $Z_i^t \in S_i$. Also, since $r_i^t$ is positive and non-decreasing either $r_i^t \to r_i$, a finite limit, or else $r_i^t \to \infty$. Consider each of the four cases separately:

Case A. Suppose $r_1^t \to r_1$ and $r_2^t \to r_2$. Since $|X|$ is bounded, it follows that $Z_i^t \to Z_i$. Therefore, $B^t \to B^1 + r_1 Z_1 - r_2 Z_2$.

Case B. Suppose $r_1^t \to r_1$ and $r_2^t \to \infty$. Since $0 < h_1 < |Z_2^t|$, it follows that $r_2^t |Z_2^t| \to \infty$. Therefore, for this case, $|B^t| \to \infty$.

Case C. Suppose $r_1^t \to \infty$ and $r_2^t \to r_2$. This is similar to Case B.

Case D. Suppose $r_1^t \to \infty$ and $r_2^t \to \infty$. For this case, $|B^t| \to \infty$, unless both the following conditions hold:

(1) $Z_1^t$ and $Z_2^t$ have the same direction in the limit and

(2) $\lim_{t \to \infty} |Z_1^t| / |Z_2^t| = \lim_{t \to \infty} r_2^t / r_1^t$

These conditions are necessary for $|B^t|$ to converge to a finite value. Suppose they hold. Let $Z^* = \lim_{t \to \infty} Z_1^t / |Z_1^t| = \lim_{t \to \infty} Z_2^t / |Z_2^t|$. Then,

$$\lim_{t \to \infty} B^t = B^1 + \lim_{t \to \infty} (r_1^t |Z_1^t| - r_2^t |Z_2^t|) Z^*.$$

For simplicity, the term $B^1$ will now be dropped. It is assumed that $B^1$ has been expressed as the difference of linear combinations of finite sets of vectors of $S_1$ and $S_2$, respectively, and that these combinations are included in $r_1^t Z_1^t$ and $r_2^t Z_2^t$, respectively.

It remains to investigate the possibility of oscillation for this case (D). Suppose the sequence $\{B^t\}$ contains a subsequence $\{B^{v_i}\}$ for which $|B^{v_i}|$ converges to a finite value. It then follows that for $j = 1, 2$, $Z_j^{v_i} / |Z_j^{v_i}| \to Z^*$, where $Z^*$ is a unit vector, $\lim r_1^{v_i} / r_2^{v_i} = \lim |Z_2^{v_i}| / |Z_1^{v_i}|$, and $B^{v_i} \to B^L$, a finite vector. Suppose $B^t$ does not converge to $B^L$. It shall be shown that this assumption leads to a contradiction with probability one. Under this assumption, the sequence $\{B^{v_i}\}$ contains a subsequence $\{B^{t_i}\}$ for which $B^{t_i} \to B^L$ and $B^{t_i + 1}$ does not converge to $B^L$. Since $Z_1^{t_i + 1} = Z_1^{t_i}(1 - e_1^{t_i} / (r_1^{t_i} + e_1^{t_i})) + (e_1^{t_i} / (r_1^{t_i} + e_1^{t_i})) X^{t_i}$,

$Z_1^{t_i+1} / |Z_1^{t_i+1}|$ converges to Z*. Similarly, $Z_2^{t_i+1} / |Z_2^{t_i+1}| \to Z^*$. Also, $\lim r_1^{t_i+1} / r_2^{t_i+1} = \lim (r_1^{t_i} + e^{t_i}) / (r_2^{t_i} + e_2^{t_i}) = \lim |Z_2^{t_i}| / |Z_1^{t_i}|$. Thus in the limit, the plane $B^{t_i+1} X = 1$ is perpendicular to the line determined by the origin and Z* and its distance from the origin oscillates finitely; the probability that this event will not occur is one unless the points which are not oriented correctly with respect to the plane $B^L X = 1$ all lie on the line determined by the origin and Z*. But in this case, it is impossible for both $r_1^t$ and $r_2^t$ to go to infinity.

From the analysis of the four cases above, it follows that either $B^t$ approaches a limit or else $|B^t| \to \infty$. It shall now be proved that if $|B^t| \to \infty$, the plane $B^t X = 1$ converges to a finite limit with probability one, if convergence is defined as follows:

Definition: Let the vector be written $B^t = c^t A^t$ where $c^t > 0$ and $|A^t| = 1$; if as $c^t \to \infty$, $A^t$ converges to a vector A, then the plane $B^t X = 1$ is said to converge to the plane $A X = 0$.

The proof below requires the following lemma.

Lemma. If $X^t \in S_1$ and $B^t X^t < -1 - o(1/t^2)$ or if $X^t \in S_2$ and $B^t X^t > 1$, then $|B^{t+1}| \leq |B^t|$, for sufficiently large t.

Proof. Suppose $X^t \in S_1$ and $B^t X^t < -1$. In this case $B^{t+1} = B^t + e_1^t X^t$. Thus $|B^{t+1}|^2 - |B^t|^2 = 2e_1^t B^t X^t + (e_1^t)^2 |X^t|^2$. Therefore, since $e_1^t = (1 - B^t X^t) / |X^t|^2 + o(1/t^2)$, $|B^{t+1}| \leq |B^t|$ if $B^t X^t \leq -1 - |X^t|^2 o(1/t^2)$. Since $|X^t|$ is bounded, the first part of the lemma follows.

Now suppose $X^t \in S_2$ and $B^t X^t > 1$. Since $B^{t+1} = B^t - e_2^t X^t$, $|B^{t+1}|^2 - |B^t|^2 = -2e_2^t B^t X^t + (e_2^t)^2 |X^t|^2$. Therefore,

since $e_2^t = (B^t X^t - 1) / |X^t|^2 + o(1/t^2)$, $|B^t + 1| \leq |B^t|$ if $B^t X^t \geq -1 + |X^t|^2 o(1/t^2)$. This inequality will hold if $B^t X^t > 1$ provided t is sufficiently large.

Since the distance of the plane $B^t X = 1$ from the origin is equal to $1/|B^t|$ it suffices to show that if the plane does not converge to a limit, no matter how close the plane gets to the origin, it will eventually move away from the origin with probability 1. If a plane which is correctly oriented with respect to some points in both $S_1$ and $S_2$ approaches the origin but does not converge to a limit, it must, when it gets sufficiently close, intersect $S_1$, $S_2$, or both. If it intersects $S_2$, since sampling is random, the probability is one that eventually a vector from $S_2$ will be sampled, which in accordance with the lemma, will result in the plane's moving away from the origin. Suppose, on the other hand, that the plane intersects $S_1$; in this case there are two possibilities:

(1) There exists no plane which passes through the origin and separates $S_1$ and $S_2$.

This implies that when t is sufficiently large, i.e., when the plane is sufficiently close to the origin, there will be vectors V of $S_1$ for which $B^t V < -1 - o(1/t^2)$; and when such a vector is sampled, the plane will move away from the origin. Furthermore, the probability that $|B^t + 1| \leq |B^t|$, i.e., the probability of a random vector not falling in the region between the parallel planes $B^t X = 1$ and $B^t X = -1 - o(1/t^2)$, becomes arbitrarily close to one as $|B^t| \to \infty$.

(2) There exist planes which pass through the origin and which separate $S_1$ from $S_2$.

In this case, if $|B^t| \to \infty$, the sequence of planes $B^t X = 1$ converges to such a plane. This can be demonstrated as follows. When t is sufficiently large no points of $S_2$ will lie between the plane $B^t X = 1$ and a separating plane which passes through the origin, i. e., $r_1^t \to \infty$ and $r_2^t \to r^* < \infty$. Let $\{B^{t_i}\}$ be a subsequence of $\{B^t\}$ for which the sequence of planes $B^{t_i} X = 1$ converges. Since $Z_1^{t_i+1} = Z_1^{t_i} (1 - e_1^{t_i} / (r_1^{t_i} + e_1^{t_i})) + (e_1^{t_i} / (r_1^{t_i} + e_1^{t_i})) X^{t_i}$, it follows that the sequence of planes $B^{t_i+1} X = 1$ converges to the limit of the sequence of planes $B^{t_i} X = 1$.

This completes the proof of the following theorem.

Theorem: If assumptions (1) through (3) hold, the plane $B^t X = 1$ converges to a limit with probability one.

5. Conclusions. It has been assumed that sampling is random from the entire population of individuals to be classified. If stratified sampling, i. e., by categories, is permitted, convergence may be made more rapid. It may be feasible, for example, to alternate categories by sampling from a given category as long as the vector sampled requires an adjustment in the estimate of the dividing plane, and as soon as a vector is obtained which is correctly oriented with respect to the plane, switching to the alternative category. Furthermore, if any information about the distribution of X is available, it might be practicable to incorporate it into the stratified sampling plan.

It is of interest to determine whether this estimating procedure is applicable to non-static situations which exhibit a shifting in the characteristics of the categories with time. It is conjectured that the answer is yes, provided that changes are sufficiently gradual and that at any given time the assumptions above hold.

## REFERENCE

(1) Papert, "Some Mathematical Theories of Learning," Fourth London Symposium on Information Theory, edited by Colin Cherry.